ProactiveBench: Benchmarking Proactiveness in Multimodal Large Language Models

Thomas De Min[1], Subhankar Roy[2], Stéphane Lathuilière[3], Elisa Ricci[1,4], and Massimiliano Mancini[1]

[1]University of Trento [2]University of Bergamo
[3]Inria, Univ. Grenoble Alpes, CNRS, LJK [4]Fondazione Bruno Kessler

thomas.demin@unitn.it

tdemin16/ProactiveBench
tdemin16/proactivebench

Abstract. Effective collaboration begins with knowing when to ask for help. For example, when trying to identify an occluded object, a human would ask someone to remove the obstruction. Can MLLMs exhibit a similar "proactive" behavior by requesting simple user interventions? To investigate this, we introduce ProactiveBench, a benchmark built from seven repurposed datasets that tests proactiveness across different tasks such as recognizing occluded objects, enhancing image quality, and interpreting coarse sketches. We evaluate 22 MLLMs on ProactiveBench, showing that (i) they generally lack proactiveness; (ii) proactiveness does not correlate with model capacity; (iii) "hinting" at proactiveness yields only marginal gains. Surprisingly, we found that conversation histories and in-context learning introduce negative biases, hindering performance. Finally, we explore a simple fine-tuning strategy based on reinforcement learning: its results suggest that proactiveness can be learned, even generalizing to unseen scenarios. We publicly release ProactiveBench as a first step toward building proactive multimodal models.

Keywords: Multimodal LLMs · Proactiveness · Benchmarking

1 Introduction

Studies in neuroscience suggest that our perception of the world arises from dynamic interaction with the environment [13,16,18,53]. Faced with incomplete or ambiguous information, we instinctively generate hypotheses, proactively search for clues, and revise our interpretations.

This ongoing cycle of inquiry and refinement is currently unexplored for multimodal large language models (MLLMs) [4,28,72], where ambiguities may arise when a user's query is unanswerable [7,66]. For instance, for the query `"What is behind the blue blocks?"` of Fig. 1, a model can answer directly by hallucinating an incorrect reply [31], or abstaining [15,64]. Such behavior is called *reactive*. Conversely, a more desirable behavior is to be *proactive* and seek additional visual cues before replying. Yet, this is complex, as a model cannot

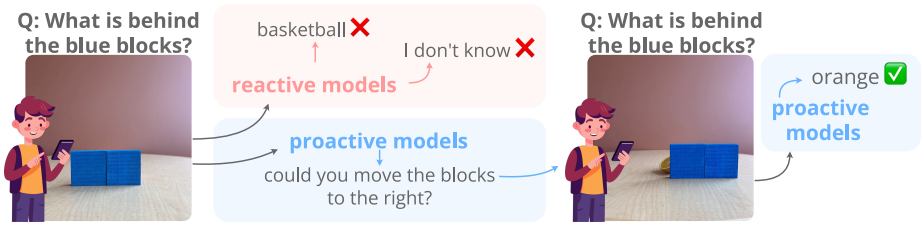


Fig. 1: Reactive *vs.* proactive models. We propose ProactiveBench, the first benchmark to evaluate MLLMs' proactiveness, *i.e.*, their ability to request additional visual cues to resolve ambiguous queries. Given an unanswerable query, a reactive model would either abstain or hallucinate. In contrast, a proactive model would ask for visual cues to disambiguate the input, enabling a correct response.

physically act in the environment. However, by recalling the previous example, the user can move the blocks to reveal the hidden object. Currently, studies focus on reactive settings, and the proactive capabilities of MLLMs are still unknown.

To fill this gap, we study whether MLLMs *can ask for help*. We introduce ProactiveBench, a novel benchmark to evaluate MLLMs' proactiveness by repurposing seven existing datasets (ROD [27], VSOD [32], MVP-N [61], ImageNet-C [17], QuickDraw [22], ChangeIt [58], and MS-COCO [33]) with different target tasks (*e.g.*, sketch recognition, product identification) that require user intervention to answer correctly. ProactiveBench captures different aspects of proactiveness: (temporal) occlusion removal, camera movement, object movement, image quality enhancement, and asking for details. Each sample has a starting ambiguous frame, a reference frame with complete information, and all the frames in between. The user intervention, guided by the model's *proactive suggestion*, produces a new frame with additional visual cues. In total, ProactiveBench contains more than 108k images grouped into 18k samples featuring 19 proactive behaviors.

We evaluate 22 state-of-the-art MLLMs (*e.g.*, LLaVA-OV [28], Qwen2.5-VL [4], InternVL3 [72]) on ProactiveBench. Our experiments suggest that models lack proactiveness, either abstaining from answering or hallucinating when visual cues are insufficient (Fig. 1). Using hints to elicit proactive behavior increases their proactive suggestion rate, but with small improvements in accuracy. Interestingly, while some MLLMs (*e.g.*, LLaVA-NeXT-Vicuna-7B, InternVL3-1B) propose more proactive suggestions than others (*e.g.*, LLaVA-OV-7B, Qwen2.5-VL-7B, InternVL3-8B), we show that this results from a lower rate of abstention on unanswerable questions, rather than a deeper understanding of the problem. Instead, conditioning on the conversation history or few-shot samples increases the proactive suggestions rate but reduces accuracy. Our results highlight that proactiveness is *not* an emerging property in MLLMs, showcasing the challenges of ProactiveBench. Additionally, we show that MLLMs can learn to be proactive through post-training with GRPO [52] equipped with tailored reward functions. Despite its simplicity, this approach yields substantial performance improvements over the original model and demonstrates strong generalization to unseen domains.

While these performance are lower than those on reference images (*e.g.*, object clearly visible, without occlusion), they suggest an interesting avenue for future works.

Contributions: (i) We formalize and explore MLLMs' proactiveness, promoting the development of models that can ask user assistance under uncertainty; (ii) We introduce ProactiveBench, an open-source benchmark to assess MLLM's proactiveness in diverse contexts; (iii) Our evaluation of 22 MLLMs on ProactiveBench reveals limited proactiveness of current models, even when explicitly hinting at being proactive, highlighting the challenges of this setting; (iv) we show that fine-tuning a model for proactiveness improves such behavior even in unseen scenarios, a promising direction toward building proactive MLLMs.

2 Related work

Benchmarking for MLLMs. While early efforts evaluate MLLMs on visual question answering [3,14,43], a second wave focused on tasks requiring reasoning and world knowledge [23,31,38,39,69]. As recent MLLMs support multiple images and videos as inputs, more complex benchmarks have been introduced to evaluate these capabilities [10,12,19,24,25,29,44,59,60]. Similarly, in the embodied AI literature, several studies evaluate LLMs [30,47,51,54,62] integrated with agents. However, none of these evaluate proactiveness to ambiguous or unanswerable queries. Related to our work, [63] and [71] show that MLLMs can perform complex tasks by actively seeking relevant information. Although both assume a collaborative setting, they focus on refining predictions by exploring modifications on a single image whose query is answerable. Liu et al. [36] explore whether MLLMs' directional guidance can support visually impaired individuals in capturing images. However, [36] limits the evaluation to a single type of proactive scenario and to single-turn conversations, not measuring the effectiveness of the MLLMs' proposed suggestions. Instead, we investigate proactiveness in seven distinct scenarios, in which actions lead to substantial changes (*e.g.*, viewpoints, quality, timestamp) over multiple turns for a single query. This enables a much more comprehensive analysis of failure cases and false proactive behaviors.

Active vision improves perception [2] by allowing an active observer to control sensing strategies (*e.g.*, viewpoint) dynamically. Active vision has been extensively studied in view planning (*i.e.*, determining optimal sensor viewpoints) [70], object recognition [5], scene and 3D shape reconstruction [56], and robotic manipulation [8]. To overcome passive systems' drawbacks, [67] introduces an open-world *synthetic* game environment in which agents actively explore their surroundings, performing multi-round abductive reasoning. Although we inherit the underlying spirit of active vision, our work differs in that: (i) ProactiveBench contains real-world images from diverse and complex scenarios; (ii) the observer receives feedback from the MLLM in natural language, fostering a collaboration of the model and the user, ideal for human-machine cooperative tasks.

3 ProactiveBench

This section introduces ProactiveBench, detailing the evaluation of MLLM proactiveness (Sec. 3.1), the benchmark creation (Sec. 3.2), and a filtering pipeline that ensures questions require MLLMs to ask for human intervention (Sec. 3.3). Model and dataset licenses are in Appendix G.

3.1 Evaluating proactiveness in MLLMs

We study MLLMs' *proactiveness*, defined as the ability to either provide a correct answer or to ask for help, suggesting actions that could make the query answerable. We evaluate proactiveness in two settings: multiple-choice question answering (MCQA) and open-ended generation (OEG).

MCQA evaluation. In this setting, models select from predefined options, allowing structured interaction with the environment and systematic assessment over multiple steps. We follow previous works on LLMs as agents [11,37] and frame the evaluation as a Markov decision process ($\mathcal{S}$, $\mathcal{A}$, π_θ, $\mathcal{R}$), over finite states space $\mathcal{S}$, discrete set of actions $\mathcal{A}$, policy π_θ (the MLLM), and reward $\mathcal{R}$. At step t, the model observes state $s_t \in \mathcal{S}$, comprising image $\mathcal{I}_t$ and valid actions $\mathcal{A}_t \subseteq \mathcal{A}$. The model selects an action a_t conditioned by question q (*e.g.*, ``what is this object?'') and state $s_t = \{\mathcal{I}_t, \mathcal{A}_t\}$, *i.e.*, $a_t \sim \pi_\theta(\cdot \mid q, s_t)$. By selecting a proactive suggestion (*e.g.*, ``move the occluding object''), state s_t transitions to s_{t+1}, leading to a new image and set of valid actions. By either abstaining (*e.g.*, ``I do not know'') or selecting a wrong category (*e.g.*, dog *vs.* cat), the evaluation stops with a wrong prediction. As environments are discrete, the policy can select proactive suggestions a finite number of times, depending on the datasets, after which the evaluation terminates with a wrong prediction. Finally, the evaluation also terminates if the model predicts the correct answer. Further implementation details are in the Appendix A.

OEG evaluation. Here, the model answers queries *without* predefined options. For this reason, evaluating OEG answers is inherently challenging as (i) they need to be interpreted and (ii) proposed actions may be inapplicable within our environments, constrained by real-world data. Therefore, to ensure fair analyses beyond such constraints, we limit the evaluation to single-turn scenarios in OEG.

Following prior works [12,35,40,41,45,48,57] we adopt an LLM-as-a-judge to score answers. In our case, the LLM is prompted to compare the answer with both proactive suggestions and category predictions, returning a binary sequence in which each bit indicates the presence (1) or absence (0) of a *valid* answer. A proactive suggestion is considered correct (*i.e.*, 1) if it is a valid mechanism to gather visual cues for the target scenario. We instruct the judge to account for variations in the answer, *e.g.*, ``change in perspective'' is accepted for ``moving the camera'', as implying the same outcome. Conversely, a proactive suggestion or category is marked as absent (*i.e.*, 0), in the answer if it is clearly missing or not valid. Due to the computational cost of open-ended generation evaluation, we limit

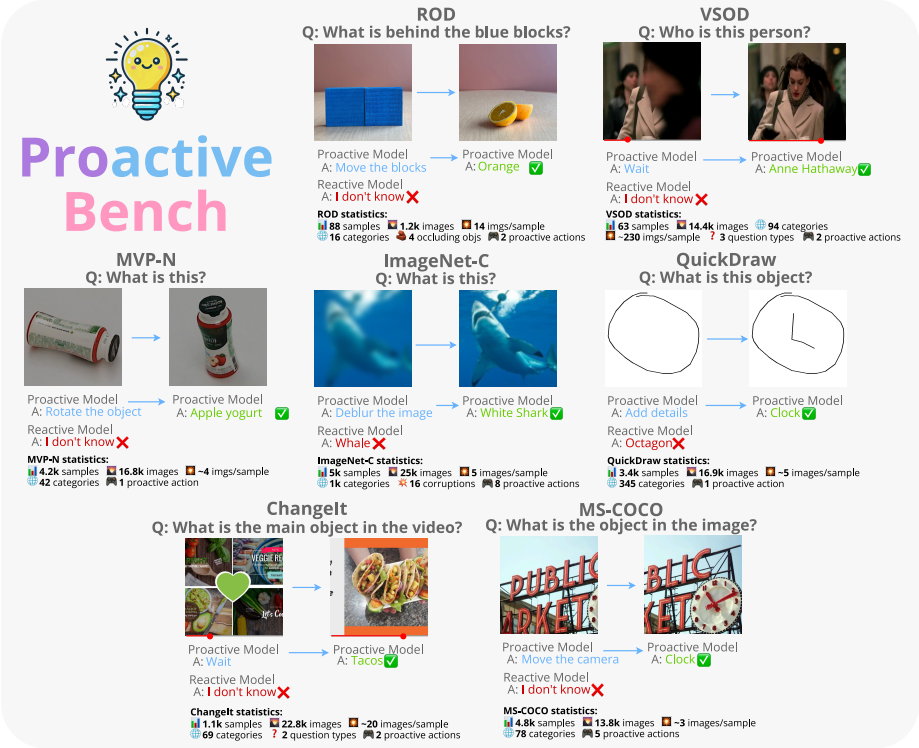


Fig. 2: ProactiveBench overview. ProactiveBench evaluates proactiveness in seven scenarios. The image shows examples of different scenarios and data statistics.

assessment to 100 examples per scenario across all scenarios of ProactiveBench. The complete LLM-as-judge prompt is provided in the Appendix B.

3.2 Benchmark construction

We introduce seven diverse scenarios to evaluate MLLMs' proactiveness. We pair each scenario with a dataset that enables multi-turn interactions through proactive suggestions in the MCQA setting. For OEG, we expand the space of valid proactive suggestions, as it is not constrained by multi-turn evaluation.

Proactive scenarios. The proposed scenarios evaluate MLLMs' in handling:

- **occluded objects** using the ROD [27] dataset, where MLLMs can ask to move the blocks to the left or right to reveal the concealed item;
- **temporal occlusions** with the VSOD [32] dataset, suggesting to inspect frames after or before the occlusion appears;
- **uninformative views** via the MVP-N [61] dataset, proposing to rotate the object or change the camera angle to help disambiguate its semantics;

- **image quality improvements** using ImageNet-C (IN-C) [17], where suggesting image quality improvements reduces the uncertainty on the content;
- **additional visual details** through QuickDraw (QD) [22], by asking the user for additional strokes, increasing the level of details in the drawing;
- **temporal ambiguities** using ChangeIt (CIT) [58], where MLLMs request past or future frames to reveal the key object or action;
- **camera movements** with MS-COCO (COCO) [33], by asking to change the point of view (*e.g.*, zoom, side movement) to better understand the scene.

An overview of the ProactiveBench scenarios is provided in Fig. 2. Additional details on each scenario are provided in Appendix A.

Annotation process. By repurposing existing datasets, we can exploit their structure and automate most of the annotation process via a rule-based procedure. For all datasets, we use their corresponding test or validation sets. For the large QD and IN-C, we sample 10 and 5 examples per category, respectively.

A challenge in creating a proactive benchmark is modeling whether a frame is informative for the target answer. In this regard, ROD, MVP-N, QD, and IN-C already provide sequences ordered from least to most recognizable frames. For example, each ROD sample has 14 frames, with the central frame being the most occluded. In earlier frames, the occluding object shifts left, revealing the target; in later frames, instead, it moves right. We therefore select the least informative frame as the initial input (*e.g.*, the first user stroke in QuickDraw). For CIT, we use the first video frame, which is typically uninformative for the task. For COCO, we select images containing a single annotated bounding box and generate challenging crops of the target object (*i.e.*, with low IoU). For VSOD, we manually identify frames where the target subject is fully occluded. Category annotations are available for all datasets except VSOD. In this case, we annotate celebrity names if they are recognized by Google Images and discard instances where recognition fails. Full dataset details are provided in Appendix A.

Note that MLLMs may still be able to recognize the target object from the least informative frames. To reduce the number of cases where proactiveness is not necessary, we employ a filtering mechanism, described in the next section.

3.3 Filtering

As most datasets are not annotated for frame informativeness (except ROD and MVP-N), some samples (*e.g.*, 55.3% in ImageNet-C) can be correctly classified from the first frame (avg. across all MLLMs). This allows models to bypass human intervention to cast correct predictions, leading to uneven performance across tasks. To focus on proactive behaviors, we filter out samples in which MLLMs can correctly guess at the first turn. Note that this filtering step removes only samples that do not contribute to estimating proactiveness, *i.e.*, in which the correct answer does not require multiple turns. Samples are filtered if they are correctly predicted at least 25% of the time in MCQA, considering all evaluated MLLMs, *during the first turn*. This strikes a good balance between removal and benchmark size. After filtering, the avg. accuracy in the first turn drops from

32.5% to 6.4%, thus requiring proactive suggestions to achieve good scores. The final benchmark counts 7,557 samples from the original size of 17,909. We further discuss the filtering details and results on unfiltered data in Appendix A.

4 Are MLLMs proactive?

This section evaluates multiple MLLMs using ProactiveBench, investigating whether they are proactive. Section 4.1 describes our evaluation protocol, tested models, and metrics used. Then, Sec. 4.2 describes ProactiveBench results, evaluating the proactiveness of several MLLMs. Finally, Sec. 4.3 reports additional ProactiveBench analysis, evaluating ways to elicit proactive suggestions.

4.1 Experimental setup

Evaluation protocol. For each evaluation step, we feed the MLLM the question, optionally a hint to elicit proactiveness, and the current image, as Sec. 3.1 describes. We additionally append the valid set of suggestions to the prompt for the MCQA setting, *i.e.*, the abstain option, proactive suggestions, and four categories, only one of which is correct (see examples in Appendix D). Hints are dataset-specific for the MCQA setting and generic for open-ended generation and lead the model towards considering proactive suggestions (*e.g.*, `‘‘Hint: rotating the object could provide a more informative view’’` for MVP-N, and for the open-ended setting `‘‘If you cannot answer this question, please tell me what I should do to help you’’`). The conversation history is always discarded unless explicitly mentioned (see Sec. 4.3). Furthermore, as VSOD and ChangeIt consist of video frames, we tell the model that the visual input is taken from a video. Finally, we rely on Qwen3-8B [68] as a judge for the open-ended generation scenario, given its reliability noted by previous work [21].

Tested models. We tested open and closed-weight MLLMs. Among open-weight models we used recent and established ones: LLaVA-1.5-7B [34], LLaVA-NeXT-7B [34] with Mistral [20] and Vicuna [6] LLMs, LLaVA-OV-0.5B, -7B, -72B [28], SmolVLM2-2.2B [42], Idefics3-8B [26], InstructBLIP [9], Qwen2.5-VL-3B, -7B, -32B, -72B [4], InternVL3-1B, -2B, -8B, -38B, -78B [72], Phi-4-Multimodal [1]. Among closed-weight models, we considered GPT-4.1, GPT-5.2, and o4-mini [46].

Metrics. Since we filtered samples that leak the answer in the first turn (Sec. 3.3), we measure meaningful proactiveness via the trajectory accuracy (*acc*), *i.e.*, the percentage of proactive sequences that lead to a correct prediction. Additionally, we also report the proactive suggestions rate (*psr*), namely the average number of human interventions requested by the model. For OEG, since the evaluation is carried out over a single turn, we separately report the percentage of answers containing the correct category (*cc*) or a valid proactive suggestion (*ps*). Note that MCQA and OEG metrics are not directly comparable, as the former measures correct answers at the end of a trajectory and the corresponding proactive suggestion rate, while the latter measures the percentage of replies that contain a correct answer or a valid proactive suggestion in the first turn.

Table 1: MCQA results on ProactiveBench. Accuracy (*acc*) and proactive suggestion rate (*psr*) of MLLMs across all ProactiveBench splits.

family	model	ROD		VSOD		MVP-N		IN-C		QD		CIT		COCO		avg.	
		acc	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*
LLaVA-1.5	7B	12.5	0.7	26.2	1.7	6.7	0.0	26.2	0.8	25.5	0.7	**44.2**	1.3	32.3	0.9	24.8	0.9
LLaVA-NeXT	Mistral-7B	0.0	0.0	0.0	0.2	1.6	0.1	10.2	0.4	1.0	0.1	17.2	1.4	1.6	0.0	4.5	0.3
	Vicuna-7B	19.3	0.7	11.9	0.5	6.5	0.1	33.2	1.3	10.2	0.9	36.6	0.9	17.1	0.3	19.3	0.7
LLaVA-OV	0.5B	44.3	2.3	9.5	1.6	12.8	0.4	24.8	1.4	**33.8**	1.5	31.1	1.4	16.9	0.4	24.8	1.3
	7B	0.0	0.0	14.3	0.4	6.7	0.0	27.8	1.0	24.3	0.4	10.4	0.3	3.2	0.0	12.4	0.3
	72B	0.0	0.0	19.0	0.4	5.0	0.1	32.2	1.2	14.3	0.2	16.9	0.5	3.7	0.0	13.0	0.3
SmolVLM2	2.2B	0.0	0.0	11.9	0.2	11.1	0.1	19.5	1.0	9.9	0.6	25.5	0.6	5.8	0.0	12.0	0.4
Idefics3	8B	31.8	1.6	19.0	2.2	7.4	0.1	32.1	1.1	12.5	0.6	12.1	0.4	9.0	0.2	17.7	0.9
InstructBLIP	7B	0.0	0.0	9.5	1.3	8.8	0.1	11.3	0.0	18.3	0.1	24.5	0.0	12.6	0.0	12.2	0.2
Qwen-2.5-VL	3B	0.0	0.0	9.5	0.0	4.9	0.0	35.9	2.0	7.9	0.2	12.4	0.3	6.3	0.0	11.0	0.4
	7B	0.0	0.0	0.0	0.0	4.3	0.0	40.5	1.3	9.9	0.1	9.8	0.1	4.9	0.0	9.9	0.2
	32B	0.0	0.0	4.8	0.0	4.6	0.0	30.9	0.4	12.3	0.0	17.4	0.4	5.5	0.0	10.8	0.1
	72B	0.0	0.0	2.4	0.2	6.7	0.0	29.2	0.9	3.1	0.1	9.3	0.3	2.0	0.0	7.5	0.2
InternVL3	1B	**61.4**	2.1	21.4	0.3	19.7	0.4	38.6	1.1	15.0	0.5	16.9	0.3	16.5	0.1	27.1	0.7
	2B	1.1	0.0	**31.0**	0.3	**20.1**	0.2	46.1	1.5	18.1	0.5	28.5	0.6	29.7	0.2	24.9	0.5
	8B	0.0	0.0	11.9	0.2	6.4	0.0	37.7	1.0	15.4	0.5	10.1	0.2	7.1	0.0	12.7	0.3
	38B	0.0	0.0	**31.0**	2.3	12.5	0.2	45.5	0.7	16.8	0.5	27.0	1.0	28.4	0.2	23.0	0.7
	78B	0.0	0.0	16.7	0.3	10.7	0.0	39.8	0.1	5.3	0.0	17.4	0.4	19.2	0.0	15.6	0.1
Phi-4-Multimodal	6B	1.1	0.0	16.7	1.0	18.9	0.0	29.8	1.6	21.9	0.4	32.6	0.6	15.2	0.2	19.4	0.5
OpenAI	o4-mini	0.0	0.0	16.7	0.6	19.8	0.0	49.0	0.2	21.6	0.0	37.9	0.8	92.8	0.0	**34.0**	0.2
	GPT-4.1	0.0	0.0	0.0	0.2	15.2	0.1	**68.2**	1.1	15.0	0.2	23.5	0.6	**94.4**	0.0	30.9	0.3
	GPT-5.2	0.0	0.0	0.0	0.2	7.8	0.1	36.6	0.3	13.6	0.1	21.7	0.5	93.6	0.0	24.8	0.2

4.2 MLLMs results in ProactiveBench

Multiple-choice question answering. Table 1 reports MLLMs' individual performance on ProactiveBench. Surprisingly, there is no simple monotonic relationship between model size and proactiveness within the tested model families, *e.g.*, InternVL3-1B outperforms InternVL3-8B in accuracy (27.1% *vs*. 12.7%) and proactive suggestions (0.7 *vs*. 0.3). Furthermore, older models (*e.g.*, LLaVA-1.5-7B) even outperform their newer and larger counterparts (*i.e.*, LLaVA-OV-72B) by a discrete margin in *acc* (24.8% *vs*. 13.0%) and *psr* (0.9 *vs*. 0.3). Interestingly, the LLM influences results, with LLaVA-NeXT Mistral achieving lower *acc* than its counterpart using Vicuna (4.5% *vs*. 19.3%). Instead, closed-source models (*e.g.*, GPT-4.1) show the best *acc*, with a low *ps*. Yet, they achieve extremely high accuracies on COCO (about 3× better than other models), suggesting potential training data contamination. Unfortunately, we cannot verify this due to the proprietary nature of the data.

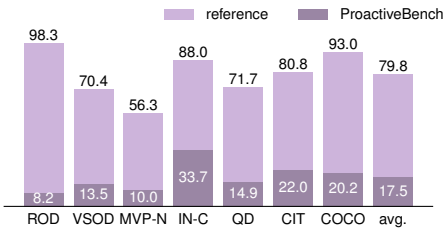


Fig. 3: Acc. in ProactiveBench *vs*. reference. Models underperform by over 60% in scenarios that require proactiveness.

To put these results in perspective, Fig. 3 compares accuracy (avg. over all models) in ProactiveBench with the *reference* setting, where we directly prompt MLLMs with the reference frame (*i.e.*, with no occlusions/ambiguity). The goal is to disentangle the recognition ability of MLLMs and their proactiveness. While MLLMs correctly

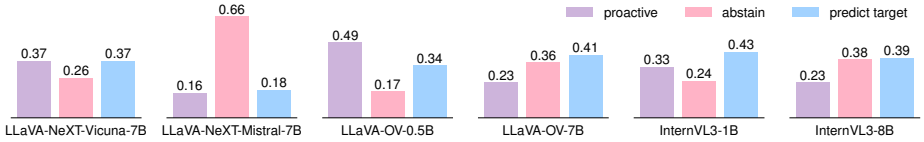


Fig. 4: Action distributions. While LLaVA-OV-7B, InternVL3-8B, and LLaVA-NeXT-Mistral-7B abstain or guess an answer, the other models prioritize proactive suggestions; thus, leveraging better visual cues and making better predictions.

Table 2: OEG results on ProactiveBench. Percentage of correct category predictions (*cc*) and valid proactive suggestions (*ps*) of MLLMs across all ProactiveBench splits.

		ROD		VSOD		MVP-N		IN-C		QD		CIT		COCO		avg.	
family	model	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*	*cc*	*ps*
LLaVA-1.5	7B	0.0	5.7	0.0	0.0	0.0	0.0	1.0	2.0	1.0	6.0	**3.0**	0.0	1.0	1.0	0.9	2.1
LLaVA-NeXT	Mistral-7B	0.0	3.4	0.0	**2.4**	0.0	0.0	1.0	34.0	**4.0**	**27.0**	2.0	6.0	4.0	1.0	1.6	10.5
	Vicuna-7B	0.0	5.7	2.4	0.0	0.0	0.0	3.0	**36.0**	0.0	23.0	**3.0**	6.0	0.0	**4.0**	1.2	**10.7**
LLaVA-OV	0.5B	0.0	1.1	0.0	0.0	0.0	0.0	0.0	11.0	2.0	2.0	1.0	0.0	0.0	1.0	0.4	2.2
	7B	0.0	1.1	0.0	**2.4**	0.0	0.0	2.0	18.0	0.0	9.0	0.0	2.0	2.0	2.0	0.6	4.9
	72B	0.0	5.7	0.0	0.0	0.0	0.0	2.0	19.0	2.0	7.0	0.0	1.0	0.0	1.0	0.6	4.8
SmolVLM2	2.2B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	3.0	0.0	0.0	0.0	0.0	1.0	0.4	0.9
Idefics3	8B	0.0	1.1	2.4	0.0	0.0	0.0	2.0	7.0	0.0	1.0	0.0	0.0	4.0	1.0	1.2	1.4
Qwen-2.5-VL	3B	0.0	1.1	2.4	0.0	0.0	0.0	2.0	22.0	2.0	3.0	0.0	1.0	3.0	0.0	1.3	3.9
	7B	0.0	**6.8**	2.4	0.0	0.0	**1.0**	3.0	31.0	0.0	16.0	1.0	8.0	2.0	3.0	1.2	9.4
	32B	0.0	3.4	0.0	0.0	0.0	0.0	2.0	25.0	0.0	3.0	2.0	**9.0**	0.0	0.0	0.6	5.8
	72B	0.0	2.3	0.0	0.0	0.0	0.0	3.0	28.0	1.0	11.0	1.0	6.0	4.0	2.0	1.3	7.0
InternVL3	1B	0.0	1.1	2.4	0.0	0.0	0.0	3.0	17.0	2.0	5.0	1.0	1.0	3.0	3.0	1.6	3.9
	2B	0.0	0.0	2.4	0.0	0.0	0.0	2.0	17.0	0.0	3.0	1.0	1.0	0.0	0.0	0.8	3.0
	8B	0.0	1.1	**4.8**	0.0	0.0	0.0	3.0	18.0	0.0	1.0	1.0	2.0	2.0	0.0	1.5	3.2
	38B	0.0	3.4	**4.8**	0.0	0.0	0.0	8.0	15.0	1.0	1.0	1.0	6.0	4.0	0.0	2.7	3.6
	78B	0.0	1.1	2.4	0.0	0.0	0.0	**11.0**	16.0	1.0	5.0	**3.0**	3.0	**7.0**	0.0	**3.5**	3.6

classify 79.8% of samples in the reference setting, they underperform by more than 60% when tasked with navigating to the correct answer through proactive suggestions. The discrepancy is quite stark in the ROD dataset, where models achieve 8.2% *acc*, while the reference counterpart reaches 98.3% on average. This demonstrates a severe lack of MLLMs' proactiveness.

We further investigate proactiveness by visualizing the action distributions, averaged across all scenarios, for proactive, abstain, and target category predictions in Fig. 4. Specifically, we compare pairs of MLLMs having different LLMs (*i.e.*, LLaVA-NeXT Mistral and Vicuna) and different parameter counts (*i.e.*, LLaVA-OV-0.5B and -7B, InternVL3-1B and -8B). While LLaVA-OV-7B, InternVL3-8B, and LLaVA-NeXT Mistral tend to abstain over sampling proactive suggestions (likely due to different training data and/or model sizes), the other three show the exact opposite behavior. Thus, they are more likely to be proactive (over 2× as likely for LLaVA-OV-0.5B) and, as a result, reach higher accuracy. A similar behavior was reported in [65], with LLaVA-NeXT Mistral abstaining more than LLaVA-NeXT Vicuna. Further results are in the Appendix E.

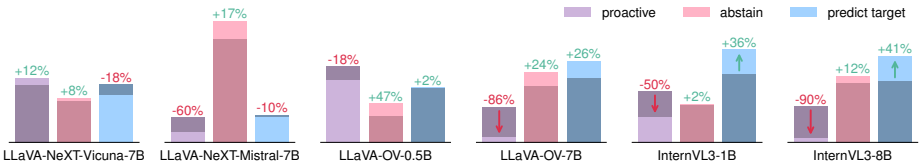


Fig. 5: Action distributions with random proactive options. Lighter bars describe variations when replacing valid proactive suggestions with invalid ones. We color-code positive and negative changes in action prob. If models still assign high prob. with random proactive actions, it implies they are not proactive and just avoid abstention.

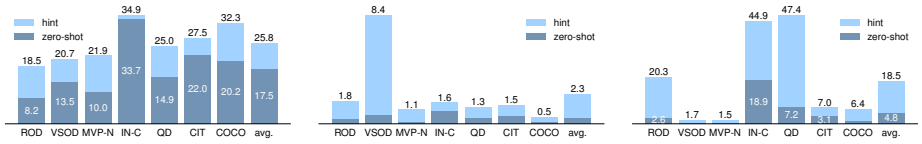
Open-ended generation. Table 2 reports MLLMs' percentage of correct first-turn predicted categories (*cc*) and valid proactive suggestions (*ps*) in OEG. Overall, even when models are not restricted to multiple-choice options, they still fail to be proactive; instead, they either abstain or hallucinate answers, much like in the MCQA setting. Similarly, there is no simple monotonic relationship between model size and proactiveness within the tested model families, suggesting that proactiveness is not a property that emerges with scale. Surprisingly, by allowing LLaVA-NeXT-Mistral to answer without constraints, it overcomes the issue with the abstention rate of Tab. 1, showing the second best accuracy in predicting valid proactive suggestions. On the other hand, InternVL3-1B superiority among open-weight models in the MCQA setting is not observed in the OEG scenario.

By examining *cc* and *ps*, we observe that the model is more accurate when proposing proactive suggestions than predicting correct categories, although their overall accuracy remains low. This behavior can be attributed to three main factors. First, we instruct the LLM-as-judge to allow for flexible rephrasings to match MLLMs' answers to the finite set of available actions, broadening the range of valid suggestions. Second, removing possible categories from the prompt increases model uncertainty, lowering categorization accuracy. Third, filtering (Sec. 3.3) makes the single-turn setting more challenging, as the model cannot rely on additional visual cues.

Overall, OEG performances are worse than in the MCQA setting. By mapping every inapplicable proactive suggestion to a random valid one, we can extend OEG to multiple turns and perform an approximate comparison between the two settings under the same metrics. The worst model in MCQA (LLaVA-NeXT-Mistral-7B) achieves 4.5% in *acc* with 0.3 *psr*. Instead, the best evaluated MLLM in multi-turn OEG (Qwen2.5-VL-7B) only achieves 3.9% *acc* with 0.1 *psr*. This confirms that MLLMs are still far from being proactive. Full results are in Appendix E.

4.3 Analyzing and eliciting MLLMs proactiveness

We now analyze MLLMs' proactiveness, investigating the influence of different prompting strategies. We focus these analyses on MCQA for its higher controllability and multi-turn evaluation, extending to OEG where possible.



(a) MCQA *acc.* **(b)** MCQA *psr.* **(c)** Open-ended gen. *ps.*

Fig. 6: Conditioning models with hints for MCQAs and OEG. Results are averaged across all MLLMs. Zero-shot refers to models not prompted with hints.

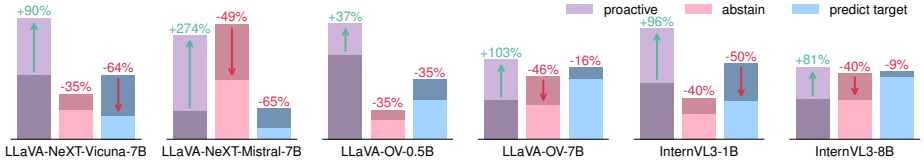
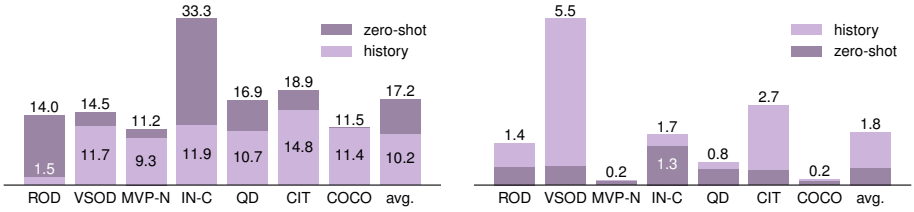


Fig. 7: Action distributions with hints. Bars describe action distributions with (light) or without (dark) hints in the prompt. We color-code positive and negative changes in action probabilities. Hinting increases the prob. of proactive suggestions.

Why some MLLMs are more likely to propose proactive suggestions? To answer this question, we replaced valid proactive suggestions with invalid ones chosen randomly from other datasets (*e.g.*, `"rewind the video"` for QuickDraw). If models that propose proactive suggestions more frequently still choose (invalid) proactive options, this implies that they are just guessing (even incorrectly) over abstaining. We limit this evaluation to the MCQA setting, as it allows for a more controlled examination. Figure 5 shows the action distribution with the same six models as Fig. 4, averaged for across all datasets. Replacing valid proactive suggestions with invalid ones substantially reduces proactive suggestion rate for LLaVA-NeXT Mistral, LLaVA-OV-7B, and InternVL3-8B (*i.e.*, −60%, −86%, and −90% relative decrease, respectively). Instead, models that predict more proactive suggestions in Fig. 4, still select proactive options in Fig. 5, even if the latter are now random and not applicable to the input scenario. LLaVA-NeXT Vicuna even increases the probability of sampling proactive suggestions (from 37% to 49%). These insights indicate that models showing a higher rate of proactive suggestions are **not** necessarily more proactive, but rather they are less prone to abstain [55], preferring unknown answers. Indeed, meaningful proactiveness occurs if and only if a model uses proactive suggestions to reach better states and improve answer accuracy. Full results are in Appendix E.

Does hinting boost proactiveness? Explicitly hinting at proactive suggestions may elicit MLLMs' proactiveness, helping to navigate to the correct answer. To evaluate this hypothesis, we add *dataset-specific* hints to MCQA prompts (*e.g.*, for ROD `"Hint: moving the occluding object might reveal what is behind it"`) and *generic* ones for the OEG (*i.e.*, `"If you cannot answer this question, please tell me what I should do to help you"`).



(a) Avg. accuracy per dataset. **(b)** Avg. proactive suggestions per dataset.

Fig. 8: Conditioning on conversation histories. Results are averaged across all MLLMs that support multi-image inference. Zero-shot refers to models not conditioned on histories. Conversation histories increase proactiveness but hurt accuracy.

We report the full list of hints used in Appendices A and B. Results are shown in Fig. 6, in terms of MCQA and OEG metrics. Extended results are in Appendix E.

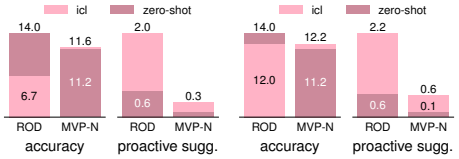
Figure 6b shows that hinting increases the *psr* in MCQA by 1.9 on average, with a significant boost in VSOD, likely due to its large exploration space. Nonetheless, the accuracy (Fig. 6a) does not surpass the random choice on average, reaching 25.8% (+8.3%) and suggesting that hinting cannot elicit proactiveness from current MLLMs. Additionally, in 16.0% of cases, MLLMs blindly choose proactive suggestions, disregarding the original task and reaching the maximum exploration steps allowed by the environment, failing to predict the correct category. Hinting also increases the percentage of valid proactive suggestions (*ps*) in OEG (Fig. 6c), with IN-C and QD showing the largest gains. Unlike other tasks that require a deeper understanding of the concept (*e.g.*, rotating the camera for MVP-N), IN-C and QD require the model to request image-quality improvements and additional details about the drawing, likely easier to interpret.

Although hinting promotes predicting proactive suggestions, models may over-exploit proactive suggestions, failing the classification even if they stumble across the reference image. Figure 7 further visualizes this by showing how action distributions change w.r.t. the six models in Fig. 4. While original distributions (darker colors) suggest that models infrequently choose proactive options, hints completely changes this behavior, with models preferring hinted actions over correct predictions, confirming their lack of proactiveness.

Does knowledge of the past elicit proactiveness? While MLLMs observe only the current state (Sec. 3.1), a key question is whether conditioning models on previous states and actions (conversation history) elicits proactiveness, *i.e.*, $\pi_\theta(\cdot \mid q, s_0, a_0, ..., s_t)$. Figure 8 shows that the average *acc* drops by 7% while the *psr* increases from 0.5 to 1.8 on average, compared to the zero-shot case, suggesting that proactiveness cannot be elicited simply through conversation histories. Although models are not explicitly "told" to be proactive, like in Fig. 6, past proactive suggestions bias models towards repeating them. MLLMs exhibited the same behavior as with "hints", repeatedly selecting proactive suggestions until reaching the maximum number of allowed steps, occurring in 12.9% of the cases. This is lower than the 16.0% observed with hints because the first action is

always unconditioned; therefore, blind selection of proactive actions occurs only if the first action is also proactive, biasing subsequent substeps.

Do few-shots improve proactiveness? We now investigate whether conditioning the policy on a few correct examples elicits proactiveness. Let $c = (q^c, s_0^c, a_0^c, ..., s_t^c, a_c^c)$ be a conversation example leading to the correct answer a_c^c. We condition the action sampling on m of such examples, $\pi_\theta(\cdot \mid c_0, ..., c_m, q, s_t)$ on ROD and MVP-N. Compared to other datasets, these two provide dense annotations indicating which frames contain a recognizable instance of the target object. This enables the automatic generation of few-shot samples composed of action sequences that transition from an ambiguous to a predictable state through proactive suggestions. We experiment with $m = 1$ and $m = 3$. Figure 9 shows how proactiveness changes with few-shot in-context learning (ICL). Compared to the base setting (zero-shot), the avg. *psr* increases by 1.4 and 0.2 on ROD and MVP-N, and 1.6 and 0.5 with one and three samples, respectively. The accuracy drops in ROD while remaining stable in MVP-N, resulting in 6.7% and 11.6% with one sample, and 12.0% and 12.2% with three. When conditioning with one sample on ROD, models either tend to predict the same category of the ICL example or blindly select proactive suggestions until reaching the maximum number of exploration steps. Scaling ICL to three examples increases *acc* in some models (*e.g.*,LLaVA-OV 7B and Phi-4-Multimodal). Generally, InternVL3-1B and LLaVA-OV-0.5B are the most prone to repeating proactive suggestions and disregarding the main task, while InternVL3-8B and SmolVLM2-2.2B tend to abstain. Similarly, in MVP-N, model errors arise either from random guesses, abstentions, or, occasionally, proactive sequences ending with incorrect predictions.



(a) 1 sample **(b)** 3 samples

Fig. 9: Conditioning on few shots. Results are averaged across all MLLMs. Zero-shot refers to models not using ICL.

5 Can MLLMs learn proactiveness from data?

In the previous section, we investigated whether proactive behavior could be elicited from MLLMs using specific conditioning strategies, but observed only marginal improvements. We now test whether proactiveness can be effectively learned and if models fine-tuned for proactiveness generalize to unseen scenarios.

Training for proactiveness. To build a training dataset that enables learning proactiveness, we follow a procedure similar to that described in Sec. 3 for constructing the MCQA setting. We train models using two scenarios, QuickDraw and COCO, because (i) they provide sufficient training data, (ii) they cover both abstract and natural images, and (iii) restricting training to a subset of datasets allows us to evaluate generalization to unseen scenarios. For COCO, we use

Table 3: Learning proactiveness. LLaVA-NeXT-Mistral-7B and Qwen2.5-VL-3B *acc* and *psr* with RL post-training. We report the original model, RL post-training with proactive reward $r_p \in \{0.5, 0.75, 1.0\}$, and the reference performance.

		in-domain				*out-of-domain*											
		QD		COCO		ROD		VSOD		MVP-N		IN-C		CIT		avg.	
model	config.	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*	*acc*	*psr*
LLaVA-NeXT Mistral-7B	original	1.0	0.1	1.6	0.0	0.0	0.0	0.0	0.2	1.6	0.1	10.2	0.4	17.2	1.4	4.5	0.3
	$r_p = 0.5$	**43.6**	1.1	57.3	1.0	**36.4**	0.9	26.2	1.7	13.7	0.2	58.6	0.6	47.2	2.2	40.4	1.1
	$r_p = 0.75$	42.6	1.1	56.6	1.0	34.1	0.8	26.2	1.7	14.1	0.2	**59.8**	0.7	42.7	2.1	39.4	1.1
	$r_p = 1.0$	11.3	4.8	**79.0**	2.6	19.3	1.1	**69.0**	9.7	**27.1**	1.1	20.9	2.8	**58.1**	4.3	**40.7**	3.7
	reference	65.6	-	95.5	-	100.0	-	57.1	-	43.6	-	88.6	-	75.3	-	75.1	-
Qwen2.5-VL-3B	original	7.9	0.2	6.3	0.0	0.0	0.0	9.5	0.0	4.9	0.0	35.9	2.0	12.4	0.3	11.0	0.4
	$r_p = 0.5$	42.9	1.2	46.1	0.6	11.4	0.1	**38.1**	0.7	**18.3**	0.1	**52.5**	2.2	52.8	1.5	37.4	0.9
	$r_p = 0.75$	**46.2**	2.0	49.4	0.8	13.6	0.2	**38.1**	0.9	17.4	0.2	50.1	2.4	**55.6**	1.7	**38.6**	1.2
	$r_p = 1.0$	0.0	5.0	**91.3**	5.2	**14.8**	11.6	2.4	60.0	0.0	3.0	5.1	2.1	6.8	18.6	17.2	15.1
	reference	65.5	-	96.0	-	100.0	-	78.6	-	51.7	-	91.5	-	84.8	-	81.2	-

its corresponding training split, while for QuickDraw, we sampled a subset that is disjoint from that used in ProactiveBench. To reduce the computational requirements and simplify optimization, we limit the training set to single-turn interactions, sampling both ambiguous and unambiguous frames. These allow the model to learn which situations require proactiveness and which direct prediction.

Knowing when to propose proactive suggestions and when to predict the correct answer requires dense annotations that we do not have and that are generally not available in standard scenarios. To overcome the need for such annotations, we train MLLMs using reinforcement learning (RL), rewarding models to jointly prioritize response efficiency (low *psr*) and *acc*. We use Group-Relative Policy Optimization (GRPO) [52], sampling 8 answers for each prompt-image pair, and rewarding each answer via the following rule: $r_c = 1$, if the answer corresponds to the correct category, $r_p \in \{0.5, 0.75, 1.0\}$, if it is a valid proactive suggestion, and $r_w = 0$ otherwise. Intuitively, if the reward for proactive suggestions is lower than correct category predictions, the model should learn to prioritize class predictions and revert to proactive suggestions when uncertain about the correct answer. Further details are in Appendix C.

Results. We compare proactiveness in MCQA pre- and post-RL for LLaVA-NeXT-Mistral-7B, the worst-performing model in the MCQA setting (see Tab. 1), and for Qwen2.5-VL-3B, one of the most widely used MLLMs. Table 3 compares MLLMs tuned with different r_p values with their original counterpart and the *reference* setting (using reference frames as in Fig. 3). Both models outperform all previously evaluated MLLMs in Tab. 1 (37.4% *vs.* 34.0% of o4-mini), except for Qwen2.5-VL-3B with $r_p = 1.0$. To this extent, setting the proactive suggestions reward lower than correct predictions, $r_p < r_c$, generally strikes a good balance between effectiveness (*e.g.*, 37.4% and 38.6% in *acc*) and efficiency (0.9 and 1.2 in *psr*). Indeed, proactive suggestions rate increases as r_p grows, *e.g.*, from 0.4 to 15.1 *psr* for Qwen2.5-VL-3B. By setting $r_p = r_c$, Qwen2.5-VL-3B excessively generates proactive suggestions, rarely producing correct predictions, lowering accuracy. Notably, we witness consistent behaviors across both seen and unseen scenarios, showing that proactiveness, once learned, generalizes to unseen domains. For

instance, CIT accuracy grows from 12.4% to 55.6% after post-training Qwen2.5-VL-3B with $r_p = 0.75$, increasing *psr* from 0.3 to 1.7, while the accuracy decreases to 5.4% when $r_p = 1.0$, reaching *psr* of 18.6.

Despite these results, the *acc* gap with the reference setting is large on average (*e.g.*, 40.7% *vs.* 75.1%), leaving many open challenges in correctly eliciting proactiveness in MLLMs. However, the generalization is encouraging and can serve as a starting point for future studies addressing this problem.

6 Conclusion

This paper presents ProactiveBench, a novel benchmark that evaluates MLLMs' proactiveness with visual inputs that require human intervention (*e.g.*, move the occluding object) to make the query answerable. ProactiveBench repurposes seven existing datasets designed for different tasks, creating sequences that evaluate proactiveness for seven distinct scenarios in single- and multi-turn interactions, both in multi-choice question answering and open-ended generation. Our findings suggest that existing MLLMs are not proactive and prefer to abstain or hallucinate. Additionally, our analysis shows that hinting at proactive suggestions increases proactiveness, but with marginal accuracy gains. Furthermore, conditioning on conversation histories and few-shot examples biases the action distribution, leading to lower accuracy. Finally, we show that while eliciting proactiveness is challenging, learning it from data is possible. We release ProactiveBench to support future works on unlocking proactive behaviors in MLLMs.

Acknowledgements

We acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support. This work is supported by the EU projects ELIAS (No.01120237) and ELLIOT (101214398). Thomas De Min is funded by NextGeneration EU. We thank the Multimedia and Human Understanding Group (MHUG) and the Fundamental AI LAB (FunAI) for their valuable feedback and insightful suggestions.

References

1. Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al.: Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv (2025)
2. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. IJCV (1988)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv (2025)
5. Browatzki, B., Tikhanoff, V., Metta, G., Bülthoff, H.H., Wallraven, C.: Active object recognition on a humanoid robot. In: ICRA (2012)

6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (2023), LMSYSOrgblog
7. Chiu, T.Y., Zhao, Y., Gurari, D.: Assessing image quality issues for real-world problems. In: CVPR (2020)
8. Chuang, I., Lee, A., Gao, D., Naddaf-Sh, M., Soltani, I., et al.: Active vision might be all you need: Exploring active vision in bimanual robotic manipulation. In: ICRA (2024)
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
10. Dingjie, S., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. In: COLM (2024)
11. Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T., Xu, K.: Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. In: NeurIPS (2024)
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: ECCV (2024)
13. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends in neurosciences (1992)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
15. Guo, Y., Jiao, F., Shen, Z., Nie, L., Kankanhalli, M.: Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. T-PAMI (2024)
16. Haskins, A.J., Mentch, J., Botch, T.L., Robertson, C.E.: Active vision in immersive, 360 real-world environments. Scientific Reports (2020)
17. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
18. Heuer, A., Ohl, S., Rolfs, M.: Memory for action: A functional view of selection in visual working memory. Visual Cognition (2020)
19. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. In: TMLR (2024)
20. Jiang, F.: Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington (2024)
21. Jiang, H., Chen, Y., Cao, Y., Lee, H.y., Tan, R.T.: Codejudgebench: Benchmarking llm-as-a-judge for coding tasks. arXiv (2025)
22. Jongejan, J., Rowley, H., Kawashima, T., Kim, J., Fox-Gieg, N.: The quick, draw!-ai experiment.(2016). QuickDraw website (2016)
23. Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., Soricut, R.: Geomverse: A systematic evaluation of large models for geometric reasoning. arXiv (2023)
24. Kazemi, M., Dikkala, N., Anand, A., Devic, P., Dasgupta, I., Liu, F., Fatemi, B., Awasthi, P., Gollapudi, S., Guo, D., et al.: Remi: A dataset for reasoning with multiple images. In: NeurIPS (2024)
25. Kil, J., Mai, Z., Lee, J., Wang, Z., Cheng, K., Wang, L., Liu, Y., Chowdhury, A., Chao, W.L.: Compbench: A comparative reasoning benchmark for multimodal llms. In: NeurIPS (2024)
26. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. arXiv (2024)

27. Lee, A.N., Bargal, S.A., Kasera, J., Sclaroff, S., Saenko, K., Ruiz, N.: Hardwiring vit patch selectivity into cnns using patch mixing. arXiv (2023)
28. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. In: TMLR (2025)
29. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
30. Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E.L., Zhang, R., et al.: Embodied agent interface: Benchmarking llms for embodied decision making. NeurIPS (2024)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023)
32. Liao, J., Duan, H., Li, X., Xu, H., Yang, Y., Cai, W., Chen, Y., Chen, L.: Occlusion detection for automatic video editing. In: ACM MM (2020)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
34. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
36. Liu, L., Yang, D., Zhong, S., Tholeti, K.S.S., Ding, L., Zhang, Y., Gilpin, L.: Right this way: Can vlms guide us to see more to answer questions? In: NeurIPS (2024)
37. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al.: Agentbench: Evaluating llms as agents. In: ICLR (2023)
38. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV (2024)
39. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences (2024)
40. Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., et al.: Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In: NeurIPS (2024)
41. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: ACL (2024)
42. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. In: COLM (2025)
43. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
44. Meng, F., Wang, J., Li, C., Lu, Q., Tian, H., Liao, J., Zhu, X., Dai, J., Qiao, Y., Luo, P., et al.: Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In: ICLR (2024)
45. Nagrani, A., Zhang, M., Mehran, R., Hornung, R., Gundavarapu, N.B., Jha, N., Myers, A., Zhou, X., Gong, B., Schmid, C., et al.: Neptune: The long orbit to benchmarking long video understanding. arXiv (2024)
46. OpenAI: Openai models. OpenAI website (2025)
47. Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., Hakkani-Tur, D.: Teach: Task-driven embodied agents that chat. In: AAAI (2022)

48. Plizzari, C., Tonioni, A., Xian, Y., Kulshrestha, A., Tombari, F.: Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In: CVPR (2025)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
51. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV (2019)
52. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv (2024)
53. Shapiro, L.: The embodied cognition research programme. Philosophy compass (2007)
54. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: CVPR (2020)
55. Shukor, M., Rame, A., Dancette, C., Cord, M.: Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. In: ICLR (2024)
56. Smith, E., Meger, D., Pineda, L., Calandra, R., Malik, J., Romero Soriano, A., Drozdzal, M.: Active 3d shape reconstruction from vision and touch. In: NeurIPS (2021)
57. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2024)
58. Souček, T., Alayrac, J.B., Miech, A., Laptev, I., Sivic, J.: Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In: CVPR (2022)
59. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR (2024)
60. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. In: NeurIPS (2024)
61. Wang, R., Wang, J., Kim, T.S., Kim, J., Lee, H.J.: Mvp-n: A dataset and benchmark for real-world multi-view object classification. In: NeurIPS (2022)
62. Wang, R., Jansen, P., Côté, M.A., Ammanabrolu, P.: Scienceworld: Is your agent smarter than a 5th grader? In: EMNLP (2022)
63. Wang, Z., Chen, C., Luo, F., Dong, Y., Zhang, Y., Xu, Y., Wang, X., Li, P., Liu, Y.: Actiview: Evaluating active perception ability for multimodal large language models. In: ACL (2025)
64. Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., Rohrbach, M.: Reliable visual question answering: Abstain rather than answer incorrectly. In: ECCV (2022)
65. Wolfe, R., Slaughter, I., Han, B., Wen, B., Yang, Y., Rosenblatt, L., Herman, B., Brown, E., Qu, Z., Weber, N., et al.: Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In: ACM-FAccT (2024)

66. Wu, T.H., Biamby, G., Chan, D., Dunlap, L., Gupta, R., Wang, X., Gonzalez, J.E., Darrell, T.: See say and segment: Teaching lmms to overcome false premises. In: CVPR (2024)
67. Xu, M., Jiang, G., Liang, W., Zhang, C., Zhu, Y.: Active reasoning in an open-world environment. In: NeurIPS (2023)
68. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv (2025)
69. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
70. Zeng, R., Wen, Y., Zhao, W., Liu, Y.J.: View planning in robot active vision: A survey of systems, algorithms, and applications. CVM (2020)
71. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: ICLR (2025)
72. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv (2025)